



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Journal Pre-proof

Machine learning-based voice assessment for the detection of positive and recovered COVID-19 patients

Carlo Robotti , Giovanni Costantini , Giovanni Saggio ,
Valerio Cesarini , Anna Calastri , Eugenia Maiorano ,
Davide Piloni , Tiziano Perrone , Umberto Sabatini ,
Virginia Valeria Ferretti , Irene Cassaniti , Fausto Baldanti ,
Andrea Gravina , Ahmed Sakib , Elena Alessi , Matteo Pascucci ,
Daniele Casali , Zakarya Zarezadeh , Vincenzo Del Zoppo ,
Antonio Pisani , Marco Benazzo



PII: S0892-1997(21)00388-X
DOI: <https://doi.org/10.1016/j.jvoice.2021.11.004>
Reference: YMVJ 3422

To appear in: *Journal of Voice*

Accepted date: 18 November 2021

Please cite this article as: Carlo Robotti , Giovanni Costantini , Giovanni Saggio , Valerio Cesarini , Anna Calastri , Eugenia Maiorano , Davide Piloni , Tiziano Perrone , Umberto Sabatini , Virginia Valeria Ferretti , Irene Cassaniti , Fausto Baldanti , Andrea Gravina , Ahmed Sakib , Elena Alessi , Matteo Pascucci , Daniele Casali , Zakarya Zarezadeh , Vincenzo Del Zoppo , Antonio Pisani , Marco Benazzo , Machine learning-based voice assessment for the detection of positive and recovered COVID-19 patients, *Journal of Voice* (2021), doi: <https://doi.org/10.1016/j.jvoice.2021.11.004>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier Inc. on behalf of The Voice Foundation.

Title

Machine learning-based voice **assessment** for the detection of positive and recovered COVID-19 patients

Authors

Carlo Robotti^{1,2,*¶}, Giovanni Costantini^{3¶}, Giovanni Saggio^{3¶}, Valerio Cesarini³, Anna Calastri¹, Eugenia Maiorano¹, Davide Piloni⁴, Tiziano Perrone⁵, Umberto Sabatini⁵, Virginia Valeria Ferretti⁶, Irene Cassaniti⁷, Fausto Baldanti^{2,7}, Andrea Gravina⁸, Ahmed Sakib⁸, Elena Alessi⁹, Matteo Pascucci⁹, Daniele Casali³, Zakarya Zarezadeh³, Vincenzo Del Zoppo³, Antonio Pisani^{10,11}, Marco Benazzo^{1,2}

Affiliations

¹ Department of Otolaryngology – Head and Neck Surgery, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

² Department of Clinical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy

³ Department of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy

⁴ Pneumology Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

⁵ Department of Internal Medicine, Fondazione IRCCS Policlinico San Matteo, University of Pavia, Pavia, Italy

⁶ Clinical Epidemiology and Biometry Unit, Fondazione IRCCS Policlinico San Matteo Foundation, Pavia, Italy

⁷ Molecular Virology Unit, Microbiology and Virology Department, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

⁸ Otorhinolaryngology Department, University of Rome Tor Vergata, Rome, Italy

⁹ Internal Medicine Unit, Ospedale dei Castelli ASL Roma 6, Ariccia, Italy

¹⁰ Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy

¹¹ IRCCS Mondino Foundation, Pavia, Italy

[†] These authors contributed equally to this work

Corresponding Author

Giovanni Saggio, PhD

Department Electronic Engineering, Roma Tor Vergata University, Rome, Italy

Address: Via del Politecnico 1, 00133 Rome, Italy

Email: saggio@uniroma2.it

Phone: (+39) 338 8690011

Declaration of interest

The authors declare the following competing interests: G.C., G.S., and A.P. are advisory members of VoiceWise S.r.l., spin-off company of University of Rome Tor Vergata (Rome, Italy) developing voice analysis solutions for diagnostic purposes; V.C. is employed by CloudWise S.r.l., company developing cloud data storage and software solutions.

Data statement

The lists of the first top 100 features obtained through the feature selection process for all tasks and all study groups are available at https://figshare.com/articles/dataset/MLVA_COVID-19/14130239.

Clinical data and audio files are not publicly available due to privacy and consent restrictions.

Moreover, data contain potentially identifying or sensitive patient information. However, they may be made available to research institutions by the authors upon reasonable request.

Abstract

Many virological tests have been implemented during the COVID-19 pandemic for diagnostic purposes, but they appear unsuitable for screening purposes. Furthermore, current screening strategies are not accurate enough to effectively curb the spread of the disease. Therefore, the present study was conducted within a controlled clinical environment to determine eventual detectable variations in the voice of COVID-19 patients, recovered and healthy subjects, and also to determine whether machine learning-based voice assessment (MLVA) can accurately discriminate between them, thus potentially serving as a more effective mass-screening tool. Three different subpopulations were consecutively recruited: positive COVID-19 patients, recovered COVID-19 patients and healthy individuals as controls. Positive patients were recruited within 10 days from nasal swab positivity. Recovery from COVID-19 was established clinically, virologically and radiologically. Healthy individuals reported no COVID-19 symptoms and yielded negative results at serological testing. All study participants provided three trials for multiple vocal tasks (sustained vowel phonation, speech, cough). **All recordings were initially divided into three different binary classifications with a feature selection, ranking and cross-validated RBF-SVM pipeline. This brought a mean accuracy of 90.24%, a mean sensitivity of 91.15%, a mean specificity of 89.13% and a mean AUC of 0.94 across all tasks and all comparisons, and outlined the sustained vowel as the most effective vocal task for COVID discrimination. Moreover, a 3-way classification was carried out on an external test set comprised of 30 subjects, 10 per class, with a mean accuracy of 80% and an accuracy of 100% for the detection of positive subjects. Within this assessment, recovered individuals proved to be the most difficult class to identify, and all the misclassified subjects were declared positive; this might be related to mid and short-term vocal traces of COVID-19, even after the clinical resolution of the infection.** In conclusion, MLVA may accurately discriminate between positive COVID-19 patients, recovered COVID-19 patients and healthy individuals. Further studies should test MLVA among larger populations and asymptomatic positive COVID-19 patients to validate

this novel screening technology and test its potential application as a potentially more effective surveillance strategy for COVID-19.

Keywords

SARS-CoV-2; machine learning; voice; cough; **recovered**; screening test; surveillance; sensitivity; accuracy.

Abbreviations

ML, machine learning; MLVA, machine learning-based voice **assessment**; NS, nasal swab; SS, serum sample

1. Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic, caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has reached more than 200 countries to date, jeopardizing healthcare systems and national administrations worldwide [1-3]. At the time of writing (September 2021), over 225 million confirmed cases and more than 4.5 million deaths had been reported globally [4]. To curb the alarming spread of the disease, several virological tests were promptly implemented, including reverse transcription-polymerase chain reaction (RT-PCR) for viral RNA detection [5], serologic testing for immunoglobulins M (IgM) and G (IgG) quantification [6] and rapid diagnostic kits [7,8]. Nevertheless, available testing strategies suffer from critical limitations in accuracy and most appropriate clinical applications [9,10]. Furthermore, inadequate testing infrastructures, high costs, lack of testing components, and long waiting times for results might have contributed to the poor control of the pandemic [11,12], leading to an underestimation of the infection's actual burden [13].

Specifically, the limitations of current screening strategies (symptoms checklists, temperature checking) [14,15] stress the need for new instruments, which should be highly sensitive, but also widely accessible, non-invasive, cost-effective, and able to provide results quickly and at scale [16]. Within this frame, the present research project was designed to test a novel COVID-19 screening tool based on voice analysis through **machine learning** (ML). Conventional voice analysis proved useful in detecting distinguishing acoustic features of pathologies impairing all structures and systems responsible for phonation, including lungs [17-19], trachea [20], larynx [21-23], vocal folds [24,25] and central nervous system [26-30]. Furthermore, encouraging results had been obtained for disorders impairing voice production mechanisms only secondarily, including cardiovascular diseases [31-34] and diabetes [35]. Advantageously, ML-based voice **assessment** (MLVA) allows processing thousands of acoustic variables simultaneously, benefiting from the computational power of properly trained algorithms [36,37]. Moreover, MLVA confirmed its efficacy even on

samples gathered through non-professional recording instruments (i.e., smartphones) [38,39], making this technology potentially available on a global scale through mobile devices.

Deep learning (DL), especially based on Convolutional Neural Networks (CNN) applied to images or Long Short Term Memory networks (LSTM), is commonly considered as the huge alternative to “traditional” machine learning pipeline methods. The advantages include the fact that it allows for the extraction of very complex features, through the repeated non-linear transformation of the data, and that it’s completely data-driven, without the requirement of a suitable pre-processing of the training set. However, deep learning is often hard to use for small datasets, requiring larger sets (often in the order of thousands and more [40]) by its nature and/or a suitable data augmentation procedure, which indeed becomes a pre-processing artifact. Moreover, DL models usually comprehend a large number of parameters and require more hardware resources for their training. Several studies [41] [42] eloquently demonstrate the possibilities of DL for pathological speech assessment, yet the datasets and the accuracy values are often comparable with traditional ML methods for the same pathologies, namely Parkinson’s disease [43] or dysphonia [44]. Moreover, a study by Hasan et al. [45] compares CNN and SVM for hyper-spectral image recognition with the results being very similar, with a slight edge of the SVM. Although apparently different from our task, hyper-spectral images are in fact translated as image inputs for CNN’s and as a set of complex features for the SVM, which is exactly how a speech-based pathology detection is carried out.

We considered the two approaches equally promising, and hugely problem-dependent. For this specific study, we chose a traditional SVM-based machine learning pipeline based on many state-of-the-art performances on pathology identification with small datasets, but also considering our need to still retain clinically relevant features in our analysis, which would be a much more difficult task using a high-abstraction DL-based approach.

Interestingly, over the last few months, some research groups have been racing worldwide to search for COVID-19 acoustic biomarkers [46], mainly focusing on cough and breath sounds and yielding promising results [47-49]. **The DiCOVA challenge within the Interspeech 2021 conference is also worth mentioning, where several teams tested algorithms for the identification of COVID-19 from crowdsourced voice samples [50], with the winning team accuracy being 87%.** Nonetheless, several limitations of these studies appear noteworthy. Firstly, data was mainly collected through web-based platforms, thus bounding researchers to rely on patients' self-declarations. Secondly, being conducted outside of controlled clinical settings, incomplete clinical data were often provided (i.e., testing type and timing, inclusion and exclusion criteria, COVID-19 symptoms). Thirdly, to the best of our knowledge, only two studies included vocal samples other than cough in the analyses [50,51]. Specifically, in our opinion, proper speech tasks could provide valuable additional features for MLVA implementation, in reason of the complex interactions between voice-production subsystems [36,52] and their peculiar impairments in COVID-19 [53-55,57]. Lastly, no ML study enrolled recovered COVID-19 patients to date, even though they may represent a crucial population for multiple reasons, including potential residual viral spreading [58-60] and severe long-term disabilities [61-63].

Therefore, the present study was designed to test MLVA as a potential screening tool for COVID-19 within a controlled clinical setting, collecting multiple vocal tasks (sustained phonation, speech, cough) with commercially available smartphones from three different subpopulations (positive COVID patients, recovered negative COVID-19 patients, healthy controls).

2. Material and methods

2.1 Study design

The study was conducted between March 2020 and October 2020 at three Italian COVID-19 units after approval of the ethics committees (IRCCS San Matteo Foundation, Pavia, Italy, reference number 20200053388; Ospedale dei Castelli, Ariccia, Italy, reference number 0064181/2020; Policlinico Tor Vergata Foundation, Rome, Italy, reference number 0012909/2020). The study was conducted following the principles stated by the Helsinki Declaration [64].

2.2 Study population

For the present study, 70 positive COVID-19 patients (group P), 70 recovered negative COVID-19 patients (group R) and 70 healthy individuals (group H) matching inclusion and exclusion criteria (Table 1) were consecutively recruited. **An additional test sample population of 10 subjects per group (15 males and 15 females; median age 53 years; IQR 37-63) was also recruited.** Positive COVID-19 patients were recruited within ten days from nasal swab (NS) positivity. For positive patients, COVID-19 pneumonia was diagnosed clinically and radiologically through chest computed tomography (CT). Recovery from COVID-19 was confirmed clinically, radiologically, and with two consecutively negative NS tests. Moreover, only recovered patients with a Lung Ultrasound (LUS) score ≤ 3 were recruited, to exclude subjects with residual pulmonary fibrosis [65-67]. Healthy individuals were recruited among hospital staff members and their acquaintances if they never had tested positive for COVID-19, reported no COVID-19 symptoms, nor had unprotected exposure to COVID-19 cases (known or suspected). Serum samples (SS) for antibodies quantification were collected from healthy participants at least 20 days after recording sessions [68-70], yielding negative results. Written informed consent was obtained from all participants. All data was pseudonymized.

2.3 Voice recordings

Recording sessions were **conducted in similar hospital rooms, with quiet environments and tolerable levels of background noise. Specifically, no machines producing static or impulsive**

noises were running in the background and no other voices were captured while recording.

Moreover, global quality of the recordings was assessed by ear by three independent audio engineers, rating samples as “acceptable” based on voice clarity, absence of noticeable reverb, absence of noticeable hiss or hum noises, and intelligible phonation. Voice samples were recorded with Huawei Y6-2019 smartphones (Huawei Technologies Co., Ltd., Shenzhen, China), in high quality and uncompressed format (.wav, 16-bit, 44.1 kHz). Devices were carefully disinfected after each use, according to the manufacturer’s instructions. Participants were instructed to sit up straight on chairs with no armrests, keeping elbows and arms relaxed to avoid arm and shoulder strain. During recording sessions, all participants removed their masks not to alter acoustic signals nor speech intelligibility [71-73]. The device’s microphone was placed 15-20 cm in front of the participants’ mouths. Three distinct vocal-tasks were performed by each participant: (1) sustained voicing of the vowel /a/ (like in “bra”), at comfortable pitch and loudness, for at least 5 seconds; (2) a common Italian saying (“a caval donato non si guarda in bocca”, literally “do not look a gift horse in the mouth”); (3) cough. Three trials were recorded for each task. For the vowel and the sentence tasks, the trial with the lowest competing noise was then selected for MLVA [74], while all three cough trials were considered for the analyses. Recordings with poor audio quality or mispronunciation errors were then discarded from the analyses (Table S1 in the Supplement). All recordings were uploaded to a secure institutional server. Audio files were then **trimmed** to retain only vocalizing sections. Each participant provided three trials for each task effortlessly. Specifically, recording sessions required no more than two minutes for each participant.

*2.4 Machine learning-based voice **assessment** (MLVA)*

MLVA was performed in five steps: **pre-processing**, feature extraction (FE), feature selection (FS), feature ranking (FR), and classification (CL). First, raw audio data of all vocal tasks underwent **pre-processing** elaboration. Specifically, Root Mean Square (RMS) normalization was applied to feed the algorithms with normalized data, thus mitigating variations related to different recording

environments. Subsequently, FE was performed embedding OpenSMILE (OpenSMILE; audEERING GmbH, Munich, Germany) [75] in a bash script, following previously validated protocols [39]. A total of 6373 unidimensional features was extracted using the configuration file of the INTERSPEECH2016 Computational Paralinguistics Challenge (IS ComParE 2016) feature dataset [75]. Subsequently, FS, FR and CL were performed using the software Weka (Waikato Environment for Knowledge Analysis; University of Waikato, Waikato, New Zealand) [76].

Starting from FS, audio files were organized into nine different datasets, with three binary comparisons between the classes, each being based on the three vocal tasks. Thus, the training set was arranged for one-versus-one comparisons, namely P versus H, P versus R and R versus H. According to a Greedy-Stepwise search method, each binary dataset underwent FS with a Correlation-Based approach [77], retaining approximately 2% of the previously extracted features. FR was performed basing on heuristic merit factors through a linear SVM classifier [78]. The first fifty top-ranking features were preserved for each dataset, retaining the most informative content while maintaining a standardized number of features. Finally, CL was conducted through the SVM classifier (Radial Basis SVM), which was selected for its effectiveness within the analysis of relatively small datasets [79,80]. **Accuracy on the training set, for each binary classifier (binary MLVA), was calculated by means of a 10-fold cross-validation, dividing the whole set into folds and using a different one of those as a validation set each time. Final accuracy is the average of the ten accuracies obtained on each 9-to-1 set. On the other hand, the test accuracy on the external set was obtained by running the pre-trained binary models on the test data, unified with a majority voting system. Since each binary model is comprised of three sub-classifiers, one per vocal task, a majority voting system was also used to unify the outputs of the three sub-classifiers. Those subjects who received three different responses from the three binary classifiers were deemed as “uncertain”. Sensitivity and specificity were obtained through confusion matrices, as well as their three-class equivalent (H Accuracy, P Accuracy, R Accuracy).**

2.5 Statistical Analysis

To compare clinical and demographic characteristics, statistical tests were performed using Stata (StataCorp 2019, Stata Statistical Software: Release 16, College Station, TX). Qualitative variables were summarized as absolute counts and percentages of each category, while quantitative variables were summarized as medians and interquartile ranges (IQRs). Fisher's exact test was used to compare categorical variables between the groups of patients. Mann-Whitney test and Kruskal-Wallis test (with Dunn's test for post-hoc comparisons) were used to compare quantitative variables between two or more groups of patients, respectively. Bonferroni's correction was applied to allow for multiple comparisons. Two-sided p-values were considered statistically significant when lower than 0.05.

3. Results

3.1 Study population

Seventy positive COVID-19 patients (group P) were consecutively enrolled at the COVID-19 Units of the enrolled institutions. To match this population, seventy participants were consecutively recruited among recovered negative COVID-19 patients and healthy individuals. **These subjects make up the training set for the binary MLVA.** Clinical and demographic data are reported in Table 2.

Moreover, thirty subjects (ten per class) have subsequently been collected for an external test sample which allowed MLVA to act as a three-way classifier.

COVID-19 symptoms were reported by 77% of both positive and recovered patients ($p > 0.90$). More than 40% of symptomatic participants of both groups reported dyspnea on exertion and asthenia ($p > 0.05$). Cough, dyspnea at rest, blocked nose, and fever were reported more frequently by positive COVID-19 patients ($p < 0.02$ in all comparisons). Contrariwise, muscle pain was reported at a higher rate by recovered subjects as a residual symptom ($p = 0.006$). Finally, no relevant differences were highlighted for the remaining screened symptoms ($p > 0.05$). At the time of enrollment, COVID-19-related pneumonia had been diagnosed clinically and radiologically in 57% of positive patients; former diagnoses of COVID-19-related pneumonia were recorded instead for 96% of recovered patients ($p < 0.001$).

3.2 Machine learning based voice *assessment* (MLVA)

Receiver Operating Characteristic (ROC) curves describing MLVA performances for each binary comparison between groups are depicted in Figure 1 and in Figures S1 and S2 in the Supplement [81]. Table 3 reports accuracy, sensitivity, specificity, and area under the ROC curve (AUC) values. Overall, **MLVA for binary classifications (on the training set)** demonstrated a mean accuracy of 90.24% (range 87.88%-92.81%), a mean sensitivity of 91.15% (range 83.58%-93.27%), a mean specificity of 89.13% (range 85.51%-92.31%) and a mean AUC of 0.94 (range 0.91-0.97) across all tasks and all comparisons. According to accuracy values, the vowel task performed as the best

discriminator within the comparison between groups P and H (90.07%) and between groups P and R (92.81%). Differently, the cough task performed as the best discriminator within the comparison between groups R and H (90.49%). Finally, radar charts highlighting the top-ranking acoustic features for all tasks and comparisons are depicted in Figure 2 and in Figures S3 to S10 in the Supplement. The lists of all top-ranking features are reported in Tables S2 to S10 in the Supplement.

For the three-way classification carried out on the external test set, accuracies of 80%, 100% and 60% have been obtained for the identification of healthy, positive and recovered subjects respectively, which brings to a mean accuracy of 80%. Noteworthy, one recovered subject was deemed as “uncertain”. For this external test, three binary classifiers were ensembled, each comprised of three ensembled sub-classifiers, one per vocal task. This unification procedure was carried out with a majority voting system. Binary accuracies for each sub-classifier were calculated only for the test subjects pertaining to the two classes considered by each binary classifier: for example, recovered subjects were not considered in evaluating the accuracy within the P versus the H classifier. Confusion matrices for each sub-classifier along with binary accuracy, sensitivity and specificity are reported in Table 4. The final confusion matrix for all the test subjects is reported in Table 5, while the compact 3x4 matrix is reported in Table 6.

4. Discussion

The present investigation, conducted within a controlled clinical setting, demonstrated that MLVA can accurately discriminate between positive COVID-19 patients, recovered negative COVID-19 patients, and healthy individuals, by detecting highly-distinguishing patterns of audio features for all tasks across all study groups. In comparison to most previous works, this study expands MLVA analyses to proper speech tasks, providing further evidence in support of the potential clinical application of this novel screening tool for COVID-19.

Binary MLVA (cross-validated on the training set) yielded promising results for all tasks and all comparisons between groups, with a satisfactory mean accuracy of 90.24% and a significantly high mean AUC of 0.94 [82]. Previous studies testing cough and breath sounds reported encouraging results in terms of accuracy (range 88.89%-98.50%) and AUC (range 0.80-0.98) [48,49]. However, comparisons with literature appear scarcely feasible, primarily for methodological reasons. Firstly, previous studies searching for COVID-19 acoustic biomarkers relied almost exclusively on cough, since it represents a well-renowned COVID-19 core symptom [83-85]. Pre-COVID-19 ML studies demonstrated the relevance of cough samples for detecting multiple respiratory conditions [86,87]. However, it is conceivable that proper speech tasks may provide additional valuable features for MLVA, potentially even more representative of the multifaceted interactions between phonatory subsystems [36,52] and their impairment in COVID-19 [53-55,57]. Indeed, the vowel task proved higher accuracy and AUC values than cough when discriminating between groups P and H and between groups P and R, demonstrating lower performances only when discriminating between groups R and H, thus confirming that speech tasks may have at least similar informative contents.

With regards to sensitivity (the ability to detect subjects with the disease) and specificity (the ability to identify healthy individuals), binary MLVA yielded satisfactory results. In particular, when discriminating between groups P and H, the vowel task demonstrated the highest sensitivity (92.11%), while the sentence task proved the highest specificity (92.31%). A preliminary observation of the selected acoustic features highlights a trend that sees domains

other than the frequency as the most prevalent. This is in line with the fact that differences in voices over the three classes are not always detectable by ear. Moreover, a prevalence of RASTA related features [88] can be observed: based on cepstral coefficients of a PLP autoregressive model, high-pass filtered in the mel-frequency domain, the RASTA domain is inherently insensitive to slowly varying spectral components, which are most often represented by background noise and differences in recording hardware and environment. On the other hand, RASTA is sensitive to eventual other background voices, which we were very careful not to include in our recordings. Pinkas [51] and Shimon [52] also obtained encouraging results through their preliminary analyses on proper speech tasks (vowel /a/, counting from 50 to 80). However, data was gathered from smaller populations of positive COVID-19 subjects and their analyses yielded lower AUC and accuracy values. Secondly, most studies gathered cough samples from crowdsourced databases, allegedly to promptly gather large datasets, nonetheless often providing incomplete medical data.

Foreseeing the need for an automatic tool, a three-way classifier was also developed by unifying the three models with a majority voting system. Thus, the features used for the classifications were the same used for the binary MLVA. This classification yielded 80% accuracy for the identification of subjects of group H, a 100% accuracy for subjects of group P and a 60% accuracy for subjects of group R, with 10% classified as “uncertain”. With a mean of 80%, the classifier still yields relevant accuracy, comparable to that of conventional nasal swabs, with the additional advantage of preliminarily discriminating recovered subjects. Furthermore, the trend of the vowel being the best discriminating task, shortly followed by the sentence, was confirmed. It is worth noting that all misclassified recovered subjects were declared positives. This might suggest that the COVID-19 “signature” persists in the voice in the mid and short-term, even when the clinical course of infection is over. This is in line with considerations by Holding et al. [89] on COVID-induced long-lasting damages to the phonatory system, which are especially concerning for voice professionals such as singers and

therefore deserve attention.

These promising results further support the potential employment of MLVA as a COVID-19 screening tool. **Regarding eventual on-site examinations, we would naturally consider the building of an uncrowded and noise-free environment very beneficial, especially regarding the complexity of the problem which requires clean datasets. However, the presence of noise-robust features like RASTA and the caveat of avoiding background voices in the recordings could be sufficient, to an extent that still needs to be thoroughly tested, for an on-site examination outside the clinical environment, such as in closed spaces placed in the territory or even in a silent room at one's house. Naturally, an automatic tool in a "real" environment will possibly require a new training process.**

The present research was conceived to overcome the critical issues of current screening strategies, such as symptoms checklists and temperature checking. Regarding symptoms checklists, a recent review concluded that commonly screened COVID-19 signs and symptoms have low diagnostic accuracy, since neither presence nor absence of symptoms is accurate enough to confirm or rule out the disease [14]. Temperature screening appears to be an unreliable, high-cost, and low-yield strategy [15]. Indeed, in a study investigating European patients' clinical features, only 45% of mild-to-moderate COVID-19 patients had fever, and the rate dropped to 9% when asymptomatic individuals were also considered [84]. Therefore, based on our preliminary results, we believe that MLVA could represent a more reliable, cost-effective, non-invasive, and widely deployable COVID-19 screening tool. Specifically, several MLVA screening scenarios could be envisioned: (a) population daily screening, potentially localizing new viral hotbeds; (b) remote testing, limiting infectious risk for healthcare workers by reducing in-person interactions; (c) alternative COVID-19 testing where virological tests are scarcely accessible or poorly available [90].

Furthermore, MLVA could also be employed at scale to preselect candidates for virological testing because of its high sensitivity and specificity. Interestingly, a recent meta-analysis highlighted that sensitivity and specificity of currently available COVID-19 diagnostics are not equally high,

ranging from 97.2% of RT-PCR analyses of sputum samples to 73.3% and even 62.3% when RT-PCR is performed on NS specimens and saliva, respectively [91]. This variability may be primarily related to COVID-19 clinical course, since chances of viral detection on biological samples depend on specific collection times [91,92]. Moreover, being time-consuming and expensive [93], these technologies appear unsuitable for reiterated population screening. Contrariwise, effective surveillance regimens (aimed at rapidly filtering infected individuals out from the population, thus preventing further spreading) should focus instead on high-frequency testing, even with lower analytic sensitivity [94]. Both high-sensitivity and low-sensitivity tests can detect the infection within its narrow transmission window, but only frequently repeated tests can spot it during its very early phases [95]. In this matter, being a low-cost and widely spreadable technology (i.e., smartphones), MLVA could potentially be employed to test large populations recurrently over time, suggesting prompt confirmation through virological diagnostics when suspected cases are detected, making this novel technology a more effective COVID-19 filter. **It is to be stressed that, in the case of quite rare diseases such as COVID-19, screening tests with high specificity and Positive Predictive value (PPV, the probability that subjects with a positive test truly have the disease) are preferable, as they offer a better “rule in” test. However, the diagnostic parameters of a screening test (such as accuracy, specificity and sensitivity) are not intrinsic properties of the test itself, but they do strongly depend upon the clinical setting in which the test is applied. Therefore, in order to reduce falsely positive results, data regarding the actual prevalence of COVID-19 should be taken into account in order to weight the results of this screening tool. Nevertheless, patients should always be sent to conventional diagnostics for confirmation (i.e., RT-PCR) in case of positive results [96,97].**

Although the test set is not large, preliminary results stresses the potential utility of MLVA as an on-site screening tool used in substitution or in addition to nasal swabs for pre-diagnosis, as well as the possibility to develop an application for real-time, remote self-assessment. In

these regards, we consider MLVA to only be a preliminary tool which should suggest a more extensive examination in case of a positive outcome.

Recruited healthy individuals had to respect strict inclusion criteria. Furthermore, SS testing was conducted at least twenty days after recording sessions, yielding negative results. Similarly, Laguarda [49] crowdsourced cough samples from healthy individuals who declared having tested negative, nevertheless not specifying testing type. For the present investigation, control subjects did not undergo baseline NS testing, since NS may yield falsely-negative results [98], especially in early phases of COVID-19 and with potentially high rates [99]. Instead, numerous studies demonstrated that seroconversion rates in positive COVID-19 patients reach almost 100% 15 to 19 days after symptoms onset [68-70], suggesting that delayed immunological confirmation might offer a more reliable strategy when recruiting healthy control subjects during the present pandemic. Noteworthy, this is the first study testing MLVA on recovered COVID-19 patients, with promising results. Specifically, the satisfactory classification accuracy obtained discriminating between positive and recovered patients suggests that MLVA may detect different COVID-19 clinical phases. Therefore, further studies should test MLVA in monitoring disease progression. Moreover, the results obtained within the discrimination between recovered and healthy individuals suggest that COVID-19 may leave detectable vocal traces even without clinically evident pulmonary impairments. In fact, all recovered patients had a Lung Ultrasound Score of 3 or lower [65-67]. The importance of recruiting recovered patients lies in the fact that these subjects might test positive again, although the reasons behind it (i.e., reinfections, new viral variants, reactivation of former infections) and eventual residual viral spreading are still debated [58-60]. Ultimately, it is expected that healthcare systems will face critical challenges in the future for the management of recovered COVID-19 patients due to potential long-term invalidating sequelae [61-63,100,101]. In this matter, MLVA could offer a feasible and low-cost strategy to detect these subjects among the general population.

Lastly, some limitations of the present investigations must be stated. Firstly, most positive and recovered COVID-19 patients (77%) presented clinical symptoms of the disease. Future studies should address this experimental approach to positive but asymptomatic patients, thus improving MLVA performances in the pre-clinical phases of COVID-19 [102]. Secondly, we are aware that our study's sample size was limited, and that the lack of sample size calculation limits the ability to draw ultimate inferences in support of a prompt employment of MLVA in clinical practice. Therefore, although promising, the results of the present study should be intended as preliminary [103]. However, the adopted rigorous methodology and the homogenous population of this study (same ethnicity, language and nationality) support the quality of our results, hopefully dispelling some skepticism towards this pioneering screening technology. Wider multicultural and multilanguage study should be designed to confirm our findings among international populations, in order to rapidly answer the pressing need for a more effective surveillance strategy for COVID-19 [104].

5. Conclusions

In conclusion, the present MLVA model demonstrated high accuracy for the discrimination between positive COVID-19 patients, recovered negative COVID-19 patients and healthy control subjects within a controlled clinical setting. **A preliminary three-way classification proves the feasibility of an automatic tool. Moreover, the prevalence of noise-robust acoustic features like the RASTA domain suggest that an on-site examination is possible, especially in sufficiently noise-free environments. Further studies should test MLVA with paucisymptomatic positive subjects, which are prevalent in the post-vaccine era, and will also focus on long-term recovered subjects. Moreover, further examinations would be beneficial especially with wider datasets among larger populations, in order to validate this novel screening instrument and answer the pressing need for a more effective surveillance strategy for COVID-19.**

6. Acknowledgements

The authors would like to thank all patients and all healthy volunteers for kindly supporting the present research project. The authors would also like to dedicate this paper to all Italian healthcare professionals strenuously fighting against COVID-19.

Journal Pre-proof

7. References

1. Tanne JH, Hayasaki E, Zastrow M, Pulla P, Smith P, Rada AG. Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. *BMJ*. 2020;368:m1090.
2. Blumenthal D, Fowler EJ, Abrams M, Collins SR. Covid-19 - Implications for the Health Care System. *N. Engl. J. Med.* 2020;383:1483-1488.
3. Açıköz Ö, Günay A. The early impact of the Covid-19 pandemic on the global and Turkish economy. *Turk. J. Med. Sci.* 2020;50:520-526.
4. World Health Organization (WHO) Coronavirus (COVID19) Dashboard. <https://covid19.who.int> (last accessed: September 15, 2021).
5. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* 2020;25:23-30.
6. Duong YT, Wright GC, Justman J. Antibody testing for coronavirus disease 2019: not ready for prime time. *BMJ*. 2020;370:m2655.
7. Bastos ML, Tavaziva G, Abidi SK, et al. Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. *BMJ*. 2020;370:m2516.
8. Guglielmi G. Fast coronavirus tests: what they can and can't do. *Nature*. 2020;585:496-498.
9. Afzal A. Molecular diagnostic technologies for COVID-19: Limitations and challenges, *J. Adv. Res.* 2020;26:149-159.
10. Jarrom D, Elston L, Washington J, et al. Effectiveness of tests to detect the presence of SARS-CoV-2 virus, and antibodies to SARS-CoV-2, to inform COVID-19 diagnosis: a rapid systematic review. *BMJ Evid. Based. Med.* 2020; 1:bmjebm-2020-111511 [Epub ahead of print].
11. Todd B. The U.S. COVID-19 Testing Failure. *Am. J. Nurs.* 2020;120:19-20.
12. Woolf SH, Chapman DA, Lee LH. COVID-19 as the Leading Cause of Death in the United States. *JAMA*. 2021;325:123-124.

13. Wu SL, Mertens A, Crider YS, et al. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nat. Commun.* 2020;11:4507.
14. Struyf T, Deeks JJ, Dinnes J, et al. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19 disease. *Cochrane Database Syst. Rev.* 2020;7:CD013665.
15. Slade DH, Sinha MS. Return to work during coronavirus disease 2019 (COVID-19): Temperature screening is no panacea. *Infect. Control. Hosp. Epidemiol.* 2020;23:1-2.
16. Irigorri N, Spackman E. Assessing the value of screening tools: reviewing the challenges and opportunities of cost-effectiveness analysis. *Public Health. Rev.* 2018;39:17.
17. Hoit JD, Solomon NP, Hixon TJ. Effect of lung volume on voice onset time (VOT). *J. Speech. Hear. Res.* 1993;36:516-520.
18. Saggio G, Bothe S. Tuberculosis Screening by Means of Speech Analysis. *CONASENSE.* 2016;1:45-56.
19. Ashraf O, Rabold E, Schlichtkrull K, et al. Voice-based screening and monitoring of chronic respiratory conditions. *Chest.* 2020;148:A1687.
20. Anderson K, Qiu Y, Whittaker AR, Lucas M. Breath sounds, asthma, and the mobile phone. *Lancet.* 2001;358:1343-1344.
21. Ezzine K, Ben Hamida A, Ben Messaoud Z, Frikha M. Towards a computer tool for automatic detection of laryngeal cancer. 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, 2016, pp. 387-392. <https://doi.org/10.1109/ATSIP.2016.7523111>.
22. Teixeira JP, Fernandes J, Teixeira F, Fernandes P. Acoustic Analysis of Chronic Laryngitis. Statistical Analysis of Sustained Speech Parameters. *BIOSIGNALS.* 2018;4:168-175.
23. Suppa A, Asci F, Saggio G, et al. Voice analysis in adductor spasmodic dysphonia: Objective diagnosis and response to botulinum toxin. *Parkinsonism Relat. Disord.* 2020;73:23-30.

24. Costa SC, Neto BGA, Fechine JM, Muppa M. Short-term cepstral analysis applied to vocal fold edema detection. *SCITEPRESS*. 2018;2:110-115.
25. Radish Kumar B, Bhat JS, Prasad N. Cepstral analysis of voice in persons with vocal nodules. *J. Voice*. 2010;24:651-653.
26. Faurholt-Jepsen M, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. *Transl. Psychiatry*. 2016;6:e856.
27. Martínez-Sánchez F, Meilán JJG, Carro J, Ivanova O. A Prototype for the Voice Analysis Diagnosis of Alzheimer's Disease. *J. Alzheimers Dis*. 2018;64:473-481.
28. Giuliano M, García-López A, Pérez S, Díaz Pérez F, Spositto O, Bossero J. Selection of voice parameters for Parkinson's disease prediction from collected mobile data. Proceedings of the twenty-second Symposium on Image, Signal Processing and Artificial Vision (STSIVA), Bucaramanga, 2019, pp. 1–3. [https://doi.org/ 10.1109/STSIVA.2019.8730219](https://doi.org/10.1109/STSIVA.2019.8730219).
29. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Trans. Biomed. Eng*. 2012;59:1264-1271.
30. Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *J. Alzheimers Dis*. 2016;49:407-422.
31. Alvear RM, Barón-López FJ, Alguacil MD, Dawid-Milner MS. Interactions between voice fundamental frequency and cardiovascular parameters. Preliminary results and physiological mechanisms. *Logoped. Phoniatr. Vocol*. 2013;38:52-58.
32. Pareek V, Sharma RK. Coronary heart disease detection from voice analysis. 2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, 2016, pp. 1–6. [https://doi.org/ 10.1109/SCEECS.2016.7509344](https://doi.org/10.1109/SCEECS.2016.7509344).
33. Maor E, Sara JD, Orbelo DM, Lerman LO, Levanon Y, Lerman A. Voice Signal Characteristics Are Independently Associated With Coronary Artery Disease. *Mayo Clin. Proc*. 2018;93:840-847.

34. Sara JDS, Maor E, Vorlaug B, et al. Non-invasive vocal biomarker is associated with pulmonary hypertension. *PLoS ONE*. 2020;15:e0231441.
35. Chitkara D, Sharma RK. Voice based detection of type 2 diabetes mellitus. 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2016, pp. 83-87.
<https://doi.org/10.1109/AEEICB.2016.7538402>.
36. Saggio G. Are Sensors and Data Processing Paving the Way to Completely Non-invasive and Not-painful Medical Tests for Widespread Screening and Diagnosis Purposes? Thirteenth International Joint Conference on Biomedical Engineering Systems and Technologies, Valletta, 2020, pp. 207-214. <https://doi.org/10.5220/0009098002070214>.
37. Saggio G, Costantini G. Worldwide Healthy Adult Voice Baseline Parameters: A Comprehensive Review. *J. Voice*. 2020. <https://doi.org/10.1016/j.jvoice.2020.08.028> [Epub ahead of print].
38. Uloza V, Padervinskis E, Vegiene A, et al. Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening. *Eur. Arch. Otorhinolaryngol*. 2015;272:3391-3399.
39. Asci F, Costantini G, Di Leo P, et al. Machine-Learning Analysis of Voice Samples Recorded through Smartphones: The Combined Effect of Ageing and Gender. *Sensors (Basel)*. 2020;20:5022.
40. Hu G, Peng X, Yang Y, Hospedales TM, Verbeek J. Frankenstein: Learning Deep Face Representations Using Small Data. *IEEE Transactions on Image Processing*. 2018;27(1):293-303.
41. Sztahó D, Gábor K, Gábor T. Deep Learning Solution for Pathological Voice Detection using LSTM-based Autoencoder Hybrid with Multi-Task Learning. In: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies. 2021;2:135-141.

42. Nissar I, Mir WA, Izharuddin, Shaikh TA. Machine Learning Approaches for Detection and Diagnosis of Parkinson's Disease - A Review. In: *2021 IEEE, 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2021;898-905.
43. Benba A, Jilbab A, Hammouch A, Sandabad S. Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease. In: *2015 International Conference on Electrical and Information Technologies (ICEIT)*. 2015;300-304.
44. Marsili L, Suppa A, Costantini G, et al. Voice Cepstral Analysis in Adductor-Type Spasmodic Dysphonia [abstract]. *Mov Disord*. 2017;32 (suppl 2).
45. Hasan H, Shafri H, Al-Habshi M. A Comparison Between Support Vector Machine (SVM) and Convolutional Neural Network (CNN) Models For Hyperspectral Image Classification. In: *IOP Conference Series: Earth and Environmental Science*. 2019;357
46. Anthes E. Alexa, do I have COVID-19? *Nature*. 2020;586:22–25.
47. Imran A, Posokhova I, Qureshi HN, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform. Med. Unlocked*. 2020;20:100378.
48. Brown C, Chauhan CJ, Grammenos A, Han J, Hasthanasombat A, Spathis D, et al. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. *arXiv*; 2020. <https://doi.org/10.1145/3394486.3412865> [Preprint].
49. Laguarda J, Hueto F, Subirana B. COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings. *IEEE OJEMB*. 2020;1:275-281.
50. Muguli A, Pinto L, Nirmala R, et al. DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics. Interspeech 2021 challenge. *arXiv:2103.09148* [preprint].
51. Pinkas G, Karny Y, Malachi A, Barkai G, Bachar G, Aharonson V. SARS-CoV-2 Detection From Voice. *IEEE Open J. Eng. Med*. 2020;1:268-274.

52. Shimon C, Shafat G, Dangoor I, Ben-Shitrit A. Artificial intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires. *J. Acoust. Soc. Am.* 2021;149(2):1120.
53. Zhang Z. Mechanics of human voice production and control. *J. Acoust. Soc. Am.* 2016;140:2614.
54. Todisco M, Alfonsi E, Arceri S, et al. Isolated bulbar palsy after SARS-CoV-2 infection. *Lancet Neurol.* 2021;20:169-170.
55. Tian S, Hu W, Niu L, Liu H, Xu H, Xiao SY. Pulmonary Pathology of Early-Phase 2019 Novel Coronavirus (COVID-19) Pneumonia in Two Patients With Lung Cancer. *J. Thorac. Oncol.* 2020;15:700-704.
56. Bai HX, Hsieh B, Xiong Z, et al. Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT. *Radiology.* 2020;296:E46-E54.
57. Fotuhi M, Mian A, Meysami S, Raji CA. Neurobiology of COVID-19. *J. Alzheimers Dis.* 2020;76:3-19.
58. Lan L, Xu D, Ye G, et al. Positive RT-PCR Test Results in Patients Recovered From COVID-19. *JAMA.* 2020;323:1502-1503.
59. Kang H, Wang Y, Tong Z, Liu X. Retest positive for SARS-CoV-2 RNA of "recovered" patients with COVID-19: Persistence, sampling issues, or re-infection? *J. Med. Virol.* 2020;92:2263-2265.
60. Stokel-Walker C. What we know about covid-19 reinfection so far. *BMJ.* 2021;372:n99.
61. Puntmann VO, Carerj ML, Wieters I, et al. Outcomes of Cardiovascular Magnetic Resonance Imaging in Patients Recently Recovered From Coronavirus Disease 2019 (COVID-19). *JAMA Cardiol.* 2020;5:1265-1273.
62. Balachandar V, Mahalaxmi I, Subramaniam M, et al. Follow-up studies in COVID-19 recovered patients - is it mandatory? *Sci. Total Environ.* 2020;729:139021.

63. Naunheim MR, Zhou AS, Puka E, et al. Laryngeal complications of COVID-19. *Laryngoscope Invest. Otolaryngol.* 2020;5:1117-1124.
64. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA.* 2013;310:2191-2194.
65. Soldati G, Smargiassi A, Inchingolo R, et al. Is There a Role for Lung Ultrasound During the COVID-19 Pandemic? *J. Ultrasound Med.* 2020;39:1459-1462.
66. Soldati G, Smargiassi A, Inchingolo R, et al. Proposal for International Standardization of the Use of Lung Ultrasound for Patients With COVID-19: A Simple, Quantitative, Reproducible Method. *J. Ultrasound Med.* 2020;39:1413-1419.
67. Perrone T, Soldati G, Padovini L, et al. A New Lung Ultrasound Protocol Able to Predict Worsening in Patients Affected by Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia. *J. Ultrasound Med.* 2021;40(8):1627-1635.
68. Long QX, Liu BZ, Deng HJ, et al. Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nat. Med.* 2020;26:845-848.
69. Zhao J, Yuan Q, Wang H, et al. Antibody Responses to SARS-CoV-2 in Patients With Novel Coronavirus Disease 2019. *Clin. Infect. Dis.* 2020;71:2027-2034.
70. Suhandynata RT, Hoffman MA, Kelner MJ, McLawhon RW, Reed SL, Fitzgerald RL. Longitudinal Monitoring of SARS-CoV-2 IgM and IgG Seropositivity to Detect COVID-19. *J. Appl. Lab. Med.* 2020;5:908-920.
71. Magee M, Lewis C, Noffs G, et al. Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *J. Acoust. Soc. Am.* 2020;148:3562.
72. Pörschmann C, Lübeck T, Arend JM. Impact of face masks on voice radiation. *J. Acoust. Soc. Am.* 2020;148:3663.
73. Muzzi E, Chermaz C, Castro V, Zaninoni M, Saksida A, Orzan E. Short report on the effects of SARS-CoV-2 face protective equipment on verbal communication. *Eur. Arch. Otorhinolaryngol.* 2021;3:1-6.

74. Lu Y, Cooke M. Speech production modifications produced by competing talkers, babble, and stationary noise. *J. Acoust. Soc. Am.* 2008;124:3261-3275.
75. Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity and Native Language. 16th Annual Conference of the International Speech Communication Association, San Francisco, 2016, pp. 2001-2005. [https://doi.org/ 10.21437/Interspeech.2016-129](https://doi.org/10.21437/Interspeech.2016-129).
76. Maimon O, Rokach L. The Data Mining and Knowledge Discovery Handbook. First edition. Springer, Berlin, 2005.
77. Ghiselli EE. Theory of Psychological Measurement. First edition. McGraw-Hill, New York, 1964.
78. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*. 2002;46:389-422.
79. Zhu Y, Ming Z, Huang Q. SVM-Based Audio Classification for Content-Based Multimedia Retrieval. In: Sebe N, Liu Y, Zhuang Y, Huang TS, editors. Multimedia Content Analysis and Mining. MCAM 2007. Lecture Notes in Computer Science, volume 4577. Springer, Berlin, 2007.
80. Costantini GM, Todisco R, Perfetti R, Basili R, Casali D. SVM based transcription system with short-term memory oriented to polyphonic piano music. 15th IEEE Mediterranean Electrotechnical Conference (MELCON), Valletta, 2010, pp. 196-201. [https://doi.org/ 10.1109/MELCON.2010.5476305](https://doi.org/10.1109/MELCON.2010.5476305).
81. Fawcett T. An introduction to ROC analysis. *Pattern Recognit. Lett.* 2006;27:861-874.
82. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1998;240:1285-1293.
83. Lechien JR, Chiesa- Estomba CM, De Siati DR, et al. Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study. *Eur. Arch. Otorhinolaryngol.* 2020;277:2251-2261.

84. Lechien JR, Chiesa-Estomba CM, Place S, et al. Clinical and epidemiological characteristics of 1420 European patients with mild-to-moderate coronavirus disease 2019. *J. Intern. Med.* 2020;288:335-344.
85. Tascini C, Sermann G, Pagotto A, et al. Blood ozonization in patients with mild to moderate COVID-19 pneumonia: a single centre experience. *Intern. Emerg. Med.* 2020;1:1-7.
86. Porter P, Abeyratne E, Swarnkar C, et al. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respir. Res.* 2019;20:81.
87. Hossain FA, Lover AA, Corey GA, Reich NG, Rahman T. FluSense: A Contactless Syndromic Surveillance Platform for Influenza-Like Illness in Hospital Waiting Areas. *IMWUT.* 2020;4:1-28.
88. Hermansky H, Morgan N. RASTA processing of speech. *Speech and Audio Processing. IEEE Transactions on.* 1994;2:578-589.
89. Holding L, Carroll TL, Nix J, Johns MM, LeBorgne WD, Meyer D. COVID-19 After Effects: Concerns for Singers. *J Voice.* 2020;S0892-1997(20)30281-2 [Epub ahead of print].
90. Kavanagh MM, Erundu NA, Tomori O, et al. Access to lifesaving medical resources for African countries: COVID-19 testing and response, ethics, and politics. *Lancet.* 2020;395:1735-1738.
91. Böger B, Fachi MM, Vilhena RO, Cobre AF, Tonin FS, Pontarolo R. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *Am. J. Infect. Control.* 2021;49:21-29.
92. Zou L, Ruan F, Huang M, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Engl. J. Med.* 2020;382:1177-1179.
93. Lauer SA, Grantz KH, Bi Q, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann. Intern. Med.* 2020;172:577-582.

94. Mina MJ, Parker R, Larremore DB. Rethinking Covid-19 Test Sensitivity - A Strategy for Containment. *N. Engl. J. Med.* 2020;383:e120.
95. Woloshin S, Patel N, Kesselheim AS. False Negative Tests for SARS-CoV-2 Infection - Challenges and Implications. *N. Engl. J. Med.* 2020;383:e38.
96. Janssens ACJW, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. *Int J Epidemiol.* 2020;49(4):1397-1403.
97. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev.* 2008;29(Suppl 1):S83-S87.
98. Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, et al. False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PLoS One.* 2020;15:e0242958.
99. Loconsole D, Passerini F, Palmieri VO, et al. Recurrence of COVID-19 after recovery: a case report from Italy. *Infection.* 2020;48:965-967.
100. Garg P, Arora U, Kumar A, Wig N. The "post-COVID" syndrome: How deep is the damage? *J. Med. Virol.* 2021;93:673-674.
101. Halpin SJ, McIvor C, Whyatt G, et al. Postdischarge symptoms and rehabilitation needs in survivors of COVID-19 infection: A cross-sectional evaluation. *J. Med. Virol.* 2021;93:1013-1022.
102. Nikolai LA, Meyer CG, Kremsner PG, Velavan TP. Asymptomatic SARS Coronavirus 2 infection: Invisible yet invincible. *Int. J. Infect. Dis.* 2020;100:112-116.
103. Topol EJ. Is my cough COVID-19? *Lancet.* 2020;396:1874.
104. Khanzada A, Hegde S, Sreeram S, et al. Challenges and Opportunities in Deploying COVID-19 Cough AI Systems. *J. Voice*, 2021. <https://doi.org/10.1016/j.jvoice.2021.08.009> [epub ahead of print].

Figures

Fig. 1. ROC curves comparing MLVA performances for all tasks within the discrimination between positive COVID-19 patients (group P) and healthy individuals (group H).

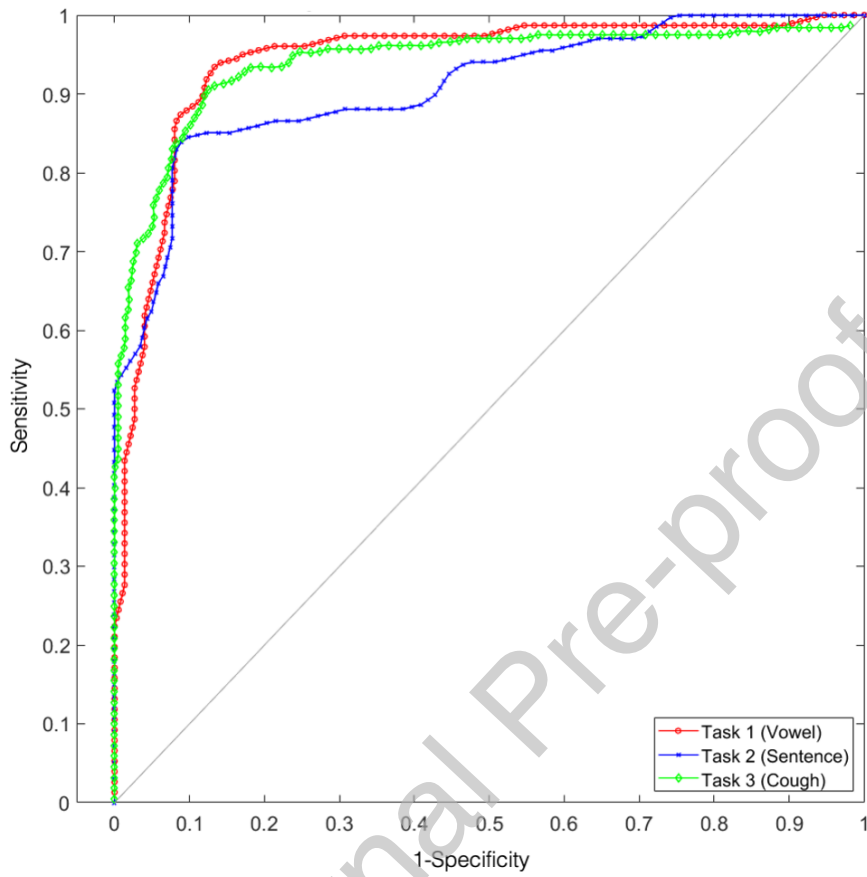
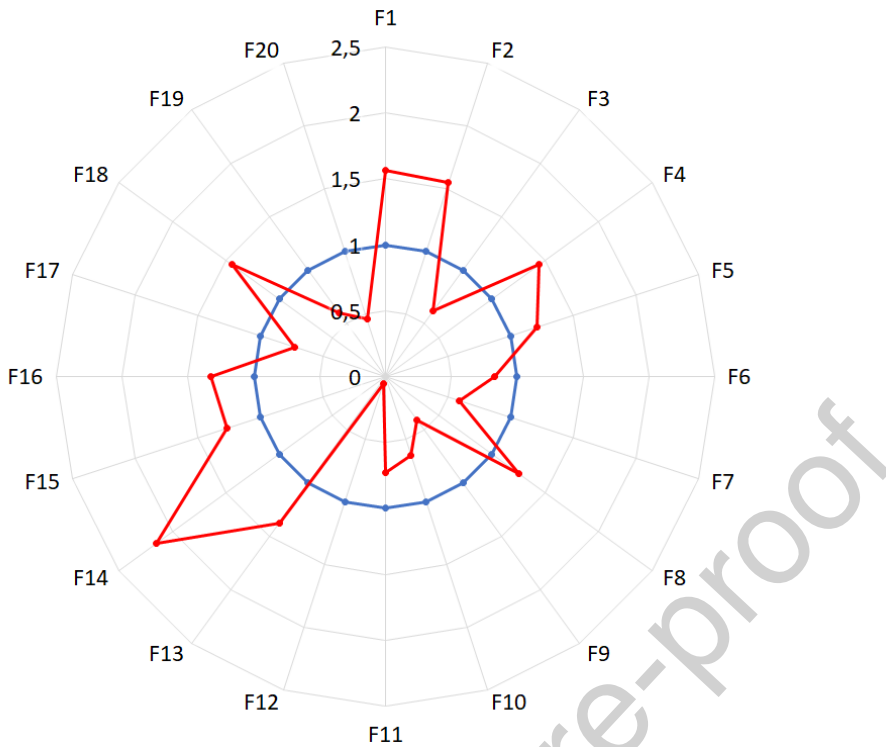


Fig. 2. Discrimination between positive COVID-19 patients and healthy individuals based on the first 20 top ranking features of the vowel task.



The red line of this radar plot corresponds to positive COVID-19 patients (group P), while the blue line corresponds to healthy individuals (group H). Each radius represents a distinct audio feature. Each point on the red line represents the feature's mean value for group P, normalized to its mean value for group H. Out of the original fifty top-ranking features, only the first twenty were reported for convenient viewing reasons. The list of all 20 top-ranking features is depicted in Table S2.

Tables

Table 1. Inclusion and exclusion criteria for the three study groups

Inclusion criteria	Group P	Group R	Group H	Exclusion criteria	Group P	Group R	Group H
Age between 18 and 80 years	■	■	■	Drugs acting on CNS	■	■	■
European ethnicity	■	■	■	Head and neck cancer	■	■	■
Italian native speaker	■	■	■	Lung cancer	■	■	■
Positive NS (< 10 days)	■	■	NA	Chemoradiation therapy	■	■	■
Two consecutive negative NS	NA	■	NA	C-PAP therapy	■	■	■
LUS ≤ 3	NA	■	NA	Tracheal intubation	■	■	■
Negative SS (> 20 days)	NA	NA	■	Tracheostomy	■	■	■

Abbreviations: P, positive COVID-19 patients; R, recovered negative COVID-19 patients; H, healthy control subjects; NS, SARS-CoV-2 nasal swab for RNA detection; LUS, lung ultrasound score; SS, SARS-CoV-2 serum sample for IgM and IgG quantification; CNS, Central Nervous System; C-PAP, Continuous Positive Airway Pressure; LUS, Lung Ultrasound score; NA, not Applicable.

Table 2. Clinical and demographic characteristics of the three study groups.

Variables	Group P (n = 70)	Group R (n = 70)	Group H (n = 70)	p-value			
				Global	P vs H	P vs R	R vs H
Age, median (IQR), years	57 (39-67)	59 (48-69)	41 (29-54)	< 0.001	< 0.001	0.215	< 0.001
Gender							
Males, n (%)	40 (57%)	45 (64%)	37 (53%)	0.402	NC	NC	NC
Females, n (%)	30 (43%)	25 (36%)	33 (47%)				
BMI, Median (IQR), kg/m²	27.8 (26.1-31.2)	26.5 (24.4-30.5)	24.3 (22.4-28.6)	0.015	0.006	0.458	0.043
Smoking habits							
Non-smokers, n (%)	35 (50%)	38 (54%)	38 (54%)	0.005	0.333	0.522	0.003
Smokers, n (%)	8 (11%)	2 (3%)	15 (21%)				
Ex-smokers, n (%)	27 (39%)	30 (43%)	17 (24%)				
COVID-19 pneumonia diagnosis, n (%)²	40 (57%)	67 (96%)	-	< 0.001	-	-	-
COVID-19 symptoms							
Presence of symptoms, n (%)	54 (77%)	54 (77%)	-	> 0.90	-	-	-
Number of symptoms, median (IQR)	2 (1-4)	2 (1-3)	-	0.096	-	-	-
Asthenia (n, %)	29 (41%)	39 (56%)	-	0.128	-	-	-
Dyspnea on exertion (n, %)	29 (41%)	31 (44%)	-	0.864	-	-	-
Cough (n, %)	34 (49%)	8 (11%)	-	< 0.001	-	-	-
Muscle pain (n, %)	10 (14%)	25 (36%)	-	0.006	-	-	-
Dysphonia (n, %)	23 (33%)	5 (7%)	-	< 0.001	-	-	-
Olfaction disorder (n, %)	13 (19%)	6 (9%)	-	0.137	-	-	-
Taste disorder (n, %)	12 (17%)	5 (7%)	-	0.119	-	-	-
Olfaction and taste disorder (n, %)	13 (19%)	6 (9%)	-	0.137	-	-	-
Dyspnea at rest (n, %)	15 (21%)	2 (3%)	-	0.001	-	-	-
Blocked nose (n, %)	11 (16%)	2 (3%)	-	0.017	-	-	-
Headache (n, %)	6 (9%)	7 (10%)	-	> 0.90	-	-	-
Fever (n, %)	7 (10%)	0 (0%)	-	0.013	-	-	-
Dysphagia (n, %)	1 (1%)	5 (7%)	-	0.209	-	-	-
Chest pain (n, %)	2 (3%)	3 (4%)	-	> 0.90	-	-	-

Data regarding COVID-19 pneumonia and COVID-19 symptoms were collected only for positive and recovered COVID-19 patients, therefore cells are left blank for healthy control subjects. Data about pneumonia for group P refer to ongoing COVID-19 pneumonia diagnosis at the time of enrollment, while for group R they refer to previously diagnosed and currently recovered COVID-19 pneumonia. Abbreviations: P, positive COVID-19 patients; R, recovered negative COVID-19 patients; H, healthy control subjects; IQR, interquartile range; NC, not calculated; BMI, body mass index; COVID-19, coronavirus disease 2019.

Table 3. Accuracy, sensitivity, specificity and Area under the curve (AUC) of Machine-Learning based Voice Analysis for all tasks and all comparisons between groups.

Comparison	Vocal task	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (CI)	Cut-off
Group P versus Group H	Vowel /a/	90.07	92.11	88.00	0.94 (0.90-0.98)	0.93
	Sentence	87.88	83.58	92.31	0.91 (0.86-0.96)	0.85
	Cough	89.44	91.28	87.50	0.92 (0.90-0.94)	0.91
Group P versus Group R	Vowel /a/	92.81	92.86	91.43	0.97 (0.95-1.00)	0.94
	Sentence	91.18	91.04	91.30	0.96 (0.92-1.00)	0.94
	Cough	91.50	93.27	89.58	0.94 (0.92-0.96)	0.92
Group R versus Group H	Vowel /a/	89.21	92.86	85.51	0.92 (0.87-0.97)	0.93
	Sentence	89.55	92.75	86.15	0.96 (0.92-0.99)	0.93
	Cough	90.49	90.63	90.36	0.92 (0.90-0.94)	0.91

Abbreviations: P, positive COVID-19 patients; R, recovered negative COVID-19 patients; H, healthy control subjects; CI, 95% confidence interval.

Table 4. Confusion matrices for each sub-classifier over the external test set, along with binary accuracy, sensitivity and specificity calculated on the two respective classes for each comparison.

#	Real Class	P versus H			P versus R			R versus H		
		Vowel /a/	Sentence	Cough	Vowel /a/	Sentence	Cough	Vowel /a/	Sentence	Cough
1	H	H	H	H	P	P	P	H	H	R
2	H	H	H	H	R	P	P	H	H	R
3	H	P	H	H	P	P	P	H	H	R
4	H	H	H	H	P	P	P	H	H	H
5	H	H	H	H	P	P	R	H	H	R
6	H	H	H	H	P	P	P	H	H	R
7	H	P	H	P	R	P	P	H	H	R
8	H	P	H	P	P	P	P	H	H	H
9	H	H	H	H	P	P	P	H	H	R
10	H	H	H	H	P	P	P	H	H	H
11	P	P	P	P	P	P	P	R	R	R
12	P	P	P	P	P	P	P	R	R	R
13	P	P	P	P	P	P	P	R	R	R
14	P	P	H	P	P	P	P	H	R	R
15	P	P	P	P	P	R	P	R	R	R
16	P	H	P	P	P	P	P	R	R	R
17	P	P	P	P	P	P	P	R	R	R
18	P	H	P	P	P	R	P	R	R	R
19	P	P	P	P	P	P	P	R	R	R
20	P	P	P	P	P	P	P	H	R	R
21	R	H	H	H	R	R	R	R	R	R
22	R	P	P	H	R	R	R	R	R	R
23	R	H	H	H	R	R	P	R	R	R
24	R	P	P	H	R	P	P	R	R	R
25	R	P	P	H	R	P	P	R	R	R
26	R	P	H	H	R	R	P	R	R	R
27	R	P	H	H	R	P	P	R	R	P
28	R	H	P	H	R	R	P	R	P	R
29	R	P	P	H	R	P	P	R	R	R
30	R	H	H	H	R	R	P	R	R	R
Accuracy (%)		75	95	90	100	80	60	100	95	60
Sensitivity (%)		80	90	100	100	100	100	100	90	90
Specificity (%)		70	100	80	100	60	20	100	100	30

Abbreviations: #, number of test subject; P, positive COVID-19 patients; R, recovered negative COVID-19 patients; H, healthy control subjects.

Table 5. Final confusion matrix for the three classifiers on the external test set along with mean accuracy and per-class accuracies.

#	Real Class	Binary Classifiers output:			Final	Error
		P versus H	P versus R	R versus H		
1	H	H	H	P	H	
2	H	H	H	P	H	
3	H	H	H	P	H	
4	H	H	H	P	H	
5	H	H	H	P	H	
6	H	H	H	P	H	
7	H	P	H	P	P	yes
8	H	P	H	P	P	yes
9	H	H	H	P	H	
10	H	H	H	P	H	
11	P	P	R	P	P	
12	P	P	R	P	P	
13	P	P	R	P	P	
14	P	P	R	P	P	
15	P	P	R	P	P	
16	P	P	R	P	P	
17	P	P	R	P	P	
18	P	P	R	P	P	
19	P	P	R	P	P	
20	P	P	R	P	P	
21	R	H	R	R	R	
22	R	P	R	R	R	
23	R	H	R	R	R	
24	R	P	R	P	P	yes
25	R	P	R	P	P	yes
26	R	P	R	R	R	
27	R	H	R	P	uncertain	uncertain
28	R	H	R	R	R	
29	R	P	R	P	P	yes
30	R	H	R	R	R	
Accuracy (%)					80	
H Accuracy (%)					80	
P Accuracy (%)					100	
R Accuracy (%)					60	

Abbreviations: #, number of test subject; P, positive COVID-19 patients; R, recovered negative COVID-19 patients; H, healthy control subjects; Final, final prediction obtained through majority voting of the three classifiers; Error, whether a mis-classification has occurred.

Table 6. Final 3x4 confusion matrix.

True class	Classified as			
	H (%)	P (%)	R (%)	Uncertain (%)
H (%)	80	20	0	0
P (%)	0	100	0	0
R (%)	0	30	60	10

Abbreviations: P, positive COVID-19 patients; R, recovered negative COVID-19 patients; H, healthy control subjects.